Denial of Information Attacks in Event Processing

Calton Pu
School of Computer Science, Georgia Institute of Technology

Extended Abstract

1. Introduction and Motivation

Automated Denial of Information Attacks. It is a common assumption in event processing that the events are "clean", i.e., they come from well-behaved and trustworthy sources. This assumption does not hold in all major open communications media for several reasons. First, adversaries may spread massive noise data, e.g., in email spam. Second, adversaries may inject potentially interesting, but obfuscating data that distracts the user's attention, e.g., in honeypot web spam pages. Third, adversaries may introduce purposefully misleading information, e.g., in phishing attacks. We call these intentional and automated attempts to generate and spread noise, obfuscating, and misleading information *denial of information* (DOI) attacks [3][6]. With the continuous advances in information technology, the quality and quantity of automatically generated DOI attacks have been increasing exponentially. In the public information area, observed DOI attacks include spam, web spam, and blog spam. Predicted DOI attacks include *spit* (spam over VoIP) and social network analysis (to be described below). The automated nature of the DOI attacks makes it increasingly difficult, if not impossible, for humans to defend themselves. The main hypothesis of this short paper is that we need to create and strengthen automated defense techniques and tools to help defuse and mitigate automated DOI attacks.

Arms Race between DOI Attacks and Defenses. Typical DOI attacks and defenses are engaged in an arms race. This can be illustrated with the co-evolution of spam messages and automated email filters employed by spam victims [1][4]. As spam messages became a serious problem, victims introduced keyword filters (Round One) to distinguish spam from legitimate messages. In response (Round Two), the spam producers adopted the misspelling attack (e.g., V1AGtRA), which is very effective against keyword filters because a victim's manual specification of keywords is quickly overrun by the automated random generation of misspellings. With the decline of keyword filters, Round Three began with the victims' adoption of statistical learning filters (e.g., Naïve Bayes), which were initially very effective. In response to learning filters (Round Four), spam producers introduced *camouflaged* content into their spam messages, which now contain both spam content and legitimate-looking camouflage designed specifically to trick the learning filters to raise their legitimacy scores. In response to camouflaged content (Round Five), victims have begun using refined learning, a process in which learning filters are incrementally trained with camouflaged spam messages. In response (Round Six), spam producers started using software tools to randomize their camouflage content to escape refined learning filters. The last 4 rounds constitute an example of the *adversarial learning* research area.

Potential Solutions for Adversarial Learning in DOI Defenses. Adversarial learning is a special kind of statistical learning, where the training data is under the influence of an adversary. The general problem of adversarial learning leads to an endless arms race, where game theory can be applied. Somewhat surprisingly, we have found an interesting solution [4] to the arms race between spam producers and victims by exploiting the asymmetry between spam messages and legitimate messages. We observed that typical legitimate messages do not contain "strong spam" tokens (e.g., misspellings of VIAGRA) that only appear in spam messages, whether they are camouflaged or not. Our main result is a training strategy (that works for Naive Bayes, SVM, and LogitBoost) that generates camouflage-resistant filters without retraining. The ongoing research applies similar ideas (and new ideas) to other adversarial

learning problems in the DOI arms race that we discuss below. These solutions that help DOI defenses "win" the arms race will bring significant breakthroughs in the area.

Deceptive Information Detection. Another major challenge posed by DOI attacks is the difficulty of distinguishing deceptive information from legitimate information, particularly when the deception is hidden in camouflage data, typically copied from previously proven legitimate information. Although the problem of human deception involves psychological and non-technical issues, automated DOI attacks present opportunities for defensive techniques that explore the static aspects of automated DOI attacks due to their programming. The solution to adversarial learning [4] illustrates this possibility.

Development of New DOI Defenses. We divide the DOI attacks into two groups: observed DOI attacks such as spam, web spam, and blogs; and expected DOI attacks such as spam through VoIP and Social Phishing [10]. As a concrete example, our experience working with observed DOI attacks such as email shows the limitations of text processing alone. In the following, we outline some existing methods for defending against DOI attacks. Then we outline an approach that combines several DOI defense mechanisms (e.g., learning filters and URL analysis) for observed DOI attacks and evaluate their efficacy. Another example of useful combination is to implement multiple DOI defense mechanisms (e.g., learning filters) further down into system and network protocol stack [8]. We also outline a second approach to develop new defenses against theoretical DOI attacks and evaluate their efficiency.

Systematic Evaluation of Automated DOI Defenses. One of the fundamental questions in automated DOI defense techniques is the evaluation method that demonstrates the superiority of that defense. For example, past evaluations of spam filters have compared their performance against manual inspection, which has limited the scale of those evaluations to a few thousands of messages. Since one of the main goals of developing automated defense mechanisms, manual evaluation methods are clearly not scalable. In comparison, we have advocated the use of large corpora (on the order of half a million messages) for more reproducible and automated evaluations [5]. We also have applied our knowledge on spam email to form a collection of about 300,000 web spam pages [2] for web spam research. We continue to contribute to the development of a systematic and reproducible evaluation method, including the publicly available large corpora as shared resources, on adversarial learning in particular and DOI research in general. Furthermore, we are applying evaluation methods inspired by scientific research to evaluate the effectiveness of DOI defenses. For example, a recent paper [1] described the initial results of evaluating the effectiveness of individual spam filtering techniques through an evolutionary study of specific spamicity tests as "genetic markers". This study covers 3 years (1/2003 through 12/2005) of spam messages, observing what tests worked (causing the spam messages containing it to go extinct) and what tests have limited effectiveness (with a significant percentage of spam messages containing it to survive).

2. An Informal Analysis of DOI Attacks and Defenses

2.1. DOI Attacks

Problem Description. We first introduced the DOI problem as an information analog of the well known denial of service (DOS) problem. We note that we are interested in DOI problems that are information-centric. For example, we leave it to the DOS researchers to handle DOS attacks on information services. As mentioned in the previous section, we are primarily interested in DOI attacks that spread massive noise, inject obfuscating and distracting data, or introduce disinformation into information sources, usually without shutting down the information service. As a concrete example, consider the most recent report (2005Q4, published in 3/2006) of the Messaging Anti-Abuse Working Group (MAAWG), which includes most of major ISPs (Internet Service Providers). The report summarizes the email processing of 127M mailboxes, with 142.5 Billion emails blocked or tagged, and 36.6 Billion emails delivered. This report shows about 80% of email traffic to be spam filtered out before they reach the destination servers.

Even with this massive filtering, most users are seeing an increasing amount of spam in their mailbox that succeeded in passing all the filters along the way.

Observed DOI Attacks. In this short paper, we focus on DOI attacks that are automatically generated, with email spam as the first example of concrete DOI attacks seen in the real world. Other growing DOI attacks include web spam, with about 20% of crawled pages with content considered to be web spam currently. Another well known growing DOI attack is blog spam, with almost all of the publicly writeable blogs being affected by some kind of automated spamming attack. Due to the automated nature of DOI attacks, effective defense mechanisms need to automate its own learning mechanism, so the defense can adapt automatically to the easy randomization incorporated into most of DOI attacks. Consequently, statistical machine learning methods have been increasingly adopted in the defense against DOI attacks. One common characteristic of these observed DOI attacks is the presence of adversarial learning, an arms race between the spammer and the victim (target) of the spam data. In the arms race, the attack and defense mechanisms co-evolve to compensate for the new capabilities acquired during the learning process. Other recent examples of DOI attacks include the work on misleading worm signature generators [7] and intrusion techniques that attempt to remain below the detection threshold of Intrusion Detection Systems.

Theoretical (Predicted) DOI Attacks. In many information flow applications we can hypothesize the possibility of DOI attacks. For example, in sensor networks, an adversary might acquire control of some sensors and make them produce misleading information. Another example of generally considered impending DOI attack is spam through VoIP (voice over IP) protocols. Anecdotally, the possibility of using software to automatically send DOI attack packets through VoIP seems quite easy and real. There are also attempts to refine some of the previously known DOI attacks. For example, email spammers have been incorporating legitimate content into spam messages in an attempt to confuse the learning filters. Another example of sophisticated spam email consists of "phishing" messages that attempt to mimic legitimate emails from a reputable company such as PayPal. One example of refined DOI attacks is the recent research at Indiana University entitled "Social Phishing", which gathered and used social network information to produce more convincing phishing messages than the current generation of "canned" phishing email, which are relatively easy to recognize through current generation of learning filters. Another interesting area of research is collaborative filtering, which is itself vulnerable to DOI attacks. One can easily imagine the adversary trying to sign on as a group and inject misleading information into the collaborative filter. Due to the apparent group effort, the confusion establishes itself relatively quickly.

2.2. Example DOI Defenses

Previous Results. Significant progress has been made in developing automated defense techniques for several concrete DOI attacks. The first example is spam filter evaluation, including a case for large corpora-based quantitative evaluation method [5], a solution for adversarial learning in resisting camouflage attacks [4], a study of spam construction technique evolution [1], and integration of diverse spam filtering techniques [3]. The second example is web spam [2]. The third example of DOI research is automated worm signature generator [7]. The fourth example is the efficient implementation of spam filters using Bloom filters [8]. The fifth example is the resistance against collusions in trust management [9]. We include a sample of these results for illustration.

Case Study of DOI Defense: Resistance to Camouflage Attacks [4]. Although learning filters such as Naïve Bayes, Support Vector Machines, and LogitBoost have been shown to be effective in classifying spam, they are vulnerable to camouflage attacks that add legitimate content/tokens to a spam message. Retraining these filters to recognize the camouflage attack tokens is only temporarily effective because new (randomized) camouflage content can still pass by the retrained filters. On surface, this "`arms race'' between camouflage spam messages and retraining of filters never ends, since the retraining is always in response to new and unpredictable camouflage content. Instead of retraining, we adopted a new design for learning filters to counter camouflage attack. The new design differs from the current generation of

learning filters (with equal number of legitimate and spam tokens used in training phase) in two ways. First, it decreases the filter sensitivity to legitimate content by limiting the number of legitimate features used in training (e.g., 25 tokens). Second, our design increases the filter sensitivity to spam content by increasing the number of spam features used in training (e.g., 9000 tokens). As a result, our new filters are able to detect spam messages with any camouflage content, without the need for retraining.

Case Study of DOI Defense: Evolution of Spam Construction [1]: We collected monthly data from SpamArchive over a three year period (from January 2003 through December 2005), accumulating more than 1.4M messages. Then, we conducted an evolutionary study by running 497 spamicity tests from SpamAssassin on each month. The population of messages testing positive for each spamicity test indicates the adoption of the spam construction technique associated with that spamicity test. This paper focuses on two evolutionary trends in our population study: extinction, where the population dwindles to zero or near zero, and co-existence, where the population maintains a consistent level or even grows, despite attempts by spamicity tests to eliminate it. We divide the factors that lead to extinction or co-existence into three groups: environmental changes, individual filtering, and collaborative filtering. We observed evidence of extinction (e.g., HTML-based obfuscation techniques), and somewhat unexpectedly, we observed evidence of co-existence between spam messages containing construction techniques and spamicity tests in filters (e.g., block list collaborative filtering).

Case Study of DOI Defense: Webb Spam Corpus [2]. Just as email spam has negatively impacted the user messaging experience, the rise of Web spam is threatening to severely degrade the quality of information on the World Wide Web. Fundamentally, Web spam is designed to pollute search engines and corrupt the user experience by driving traffic to particular spammed Web pages, regardless of the merits of those pages. We identify an interesting link between email spam and Web spam, and we use this link to demonstrate a novel technique for extracting large Web spam samples from the Web. Then, we present the Webb Spam Corpus (a first-of-its-kind, large-scale, and publicly available Web spam data set that was created using our automated Web spam collection method. The corpus consists of nearly 350,000 Web spam pages, making it more than two orders of magnitude larger than any other previously cited Web spam data set. Finally, we identify several application areas where the Webb Spam Corpus may be especially helpful. Interestingly, since the Webb Spam Corpus bridges the worlds of email spam and Web spam, we note that it can be used to aid traditional email spam classification algorithms through an analysis of the characteristics of the Web pages referenced by email messages.

Case Study of DOI Attack: Misleading Worm Signature Generators [7]: Several syntactic-based automatic worm signature generators, e.g., Polygraph, have recently been proposed. These systems typically assume that a set of suspicious flows are provided by a flow classifier, e.g., a honeynet or an intrusion detection system, that often introduces “noise” due to difficulties and imprecision in flow classification. The algorithms for extracting the worm signatures from the flow data are designed to cope with the noise. It has been reported that these systems can handle a fairly high noise level, e.g., 80% for Polygraph. In this paper, we show that if noise is introduced deliberately to mislead a worm signature generator, a much lower noise level, e.g., 50%, can already prevent the system from reliably generating useful worm signatures. We describe a new and general class of attacks whereby a worm can combine polymorphism and misleading behavior to intentionally pollute the dataset of suspicious flows during its propagation and successfully mislead the automatic signature generation process. This study suggests that unless an accurate and robust flow classification process is in place, automatic syntactic-based signature generators are vulnerable to such noise injection attacks.

Case Study of DOI Defense: TrustGuard [9]. Reputation systems have been popular in estimating the trustworthiness and predicting the future behavior of nodes in a large-scale distributed system where nodes may transact with one another without prior knowledge or experience. One of the fundamental challenges in distributed reputation management is to understand vulnerabilities and develop mechanisms that can minimize the potential damages to a system by malicious nodes. We identify three vulnerabilities that are detrimental to decentralized reputation management and propose the TrustGuard safeguard framework. First, we provide a dependable trust model and a set of formal methods to handle

strategic malicious nodes that continuously change their behavior to gain unfair advantages in the system. Second, a transaction based reputation system handles malicious nodes that produce flooding feedbacks with fake transactions. Third, we filter out dishonest feedback inserted by malicious nodes through collusion. Our experiments show that, comparing with existing reputation systems, our framework is highly dependable and effective in countering malicious nodes regarding strategic oscillating behavior, flooding malevolent feedbacks with fake transactions, and dishonest feedbacks.

3. Approaches to DOI Defenses

Analysis of Automated Defenses for Observed DOI Attacks. For short term deliverables, we will continue to develop DOI defense techniques for observed DOI attacks such as spam, web spam, blog, automated worm signature generators, and efficient spam filter implementation. Some of these efforts will build on and complement the research results summarized under the "Case Study of DOI" label above. For example, the combination of multiple spam filtering techniques to improve the efficiency of spam filtering can produce quick and positive results. Another example is the combination of trust management with collaborative filtering, since we need to improve the precision of collaborative filtering through the discovery of colluding parties, just as the web spam researchers have found. Some other efforts in this thrust will apply known techniques to new areas, similar to applying email spam filtering results to distinguish web spam pages. An example of a new effort in this area would apply email spam and web spam results to combat automated blog spam.

Automated Defenses for Theoretical DOI Attacks. To achieve revolutionary advances in this research, we will develop new DOI defense techniques for predicted DOI attacks that may not have become popular. Given our reliance on large public corpora collected from real attacks (see the Evaluation section), it is non-trivial to demonstrate the effectiveness of our defense mechanisms quantitatively. We are designing new defense mechanisms for these predicted DOI attack areas in coordination with the development of synthetic corpora that contain the predicted attacks such as spam over VoIP and social phishing. Another interesting example of theoretical DOI attacks is the targeted false positive attack, where targeted legitimate content (e.g., official email from a competing company) is sent with strong spam tokens (a reverse camouflage), so the targeted content become tainted in many spam filters by association through the normal incremental learning process. Consequently, the targeted content will receive high spamicity scores and be filtered out in many stages of email transmission. This attack is particularly important for intelligence applications, since it is useful in the hiding of targeted content from human users. We will apply our experience and results from Research Thrust 1 on observed DOI attacks, including the quantity and variety of corpora content needed for reproducible experiments.

Example Applications. An important task in mission-critical enterprise applications is the continuous integration of new information from various information sources. This task is particularly important and challenging for sensors that generate up-to-date information about the real world (called *live information*). Live information is characterized by three properties. First, live information is new information originated from sensors connected either to the real world (e.g., real-time sensor data) or to an artificial world (e.g., simulations or human analysis). In both cases, the sensors detect situations or measure variables that may lead to complex events that are difficult to predict, e.g., indications of an impending terrorist attack. Second, live information is perishable, with value that decays with time, e.g., the information on a planned terrorist attack becomes less valuable after the attack has occurred. Third, live information must be delivered in ways that preserve broad metrics of quality, including consistency, reliability, and security. For mission-critical applications, reliability and security are paramount. A demonstration application, Live Information Integration, consists of an information flow connecting many components, starting from sensors that capture and transmit new information, through intermediate selection, filtering, and combination processes, to real-time event detection programs and backend database servers that store the information for further analysis. This application is highly dynamic,

evolving with many sensors joining and leaving the system. It also has a highly variable workload, since the information volume gathered by the sensors varies according to external stimulus.

Example Application of DOI Defenses. Live information sources are inherently vulnerable to DOI attacks. For example, sensor networks can easily allow DOI attacks when the adversary hijacks some of the sensors, making them produce noise, obfuscating or distracting information, or disinformation. Our hypothesis is that currently known DOI defense techniques and tools may be effective in related application areas. For example, the solution to adversarial learning in camouflaged email spam may be applicable to distinguish camouflage sources (that produce misleading data in addition to legitimate data) from legitimate sources in other areas such as sensors, in addition to the identification of phishing messages and social phishing messages. Although these are clearly very difficult problems if we only applied text filtering, we may be able to find better solutions through the integration of multiple spam filtering techniques [3] that combines statistical learning filters and semantic interpretation of email content, e.g., URLs.

4. Evaluation Approach

Systematic Evaluation of Automated DOI Defenses. One of the fundamental questions in automated DOI defense techniques is the evaluation method that demonstrates the superiority of that defense. For example, past evaluations of spam filters have compared their performance against manual inspection. Although manual inspection of emails is considered the "golden standard" without errors, it has inherent limitations (of up to a few thousands of messages) on the scale of those experiments to the evaluations. In comparison, we have advocated the use of large corpora (on the order of half a million messages) since we have found that evaluations done with smaller corpora yield large variances in experimental results [5]. Using published small corpora, repeating the same experiments results in false positives ranging from 0% to 46% and false negatives ranging from 3% to 96%. For researchers in traditional machine learning areas such as speech recognition, the need for large corpora used in reproducible and comparable experiments would be unsurprising, since they have reached consensus for such need in early 90's.

Evaluation Method and Infrastructure. The large corpora used in our email experiments, the SpamArchive for confirmed spam messages and the Enron Corpus for legitimate messages, are in public domain. However, our study on the evolution of spam construction techniques [1] shows the need for maintaining updated large corpora for consistent evaluations of spam filters. Recent glitches in the volunteer-supported SpamArchive collection resulted in gaps during early 2006, showing the need for alternative public large collections. We have started such an effort in the collection of a large web spam corpus of about 350,000 web spam pages [2], the first of its kind to the best of our knowledge. In this work, we applied our experience and knowledge on spam email filtering to the web spam area. By assuming a connection between spam messages and web spam, we used URLs contained in spam messages to accumulate a corpus of web spam documents to support future web spam research that can evaluate their algorithms on published large corpora. This research will continue to contribute to the development of a systematic and reproducible evaluation method, including the publicly available large corpora as shared resources, on adversarial learning in particular and DOI research in general.

Evaluation of Theoretical DOI Attacks. Part of the revolutionary advance in this research is the development of infrastructure and methods to evaluate the defense mechanisms developed for theoretical (predicted) DOI attacks such as spam over VoIP, targeted false positive attacks, and social network analysis attacks mentioned above. We anticipate the need for evaluation methods beyond traditional evaluation methods based on precision and recall. For example, our evolutionary study [1] shows some unexpected results of an observation-based study of spam construction techniques that is orthogonal to statistical learning. For example, collaborative filtering has been able to identify known "bad" URLs and put them into blacklists such as WS.SURBL.ORG, about 60% of spam messages continue to carry such URLs, despite their presence in the blacklists. There are two efforts on the evaluation of defense

mechanism to handle theoretical DOI attacks predicted by our research. One is on the construction of reasonable collections for evaluation before the real attacks start, a capability that would allow us to evaluate the defense mechanisms effectively and preemptively. The other is the design of novel evaluation methods such as the evolutionary study [1].

REFERENCES

[1] C. Pu and S. Webb, "Observed Trends in Spam Construction Techniques: A Case Study of Spam Evolution." To appear in the Proceedings of the *2006 Conference on Email and Anti-Spam* (CEAS'06), Mountain View, CA, July 2006.

[2] S. Webb, J. Caverlee, and C. Pu, "Introducing the Webb Spam Corpus: Using Email Spam to Identify Web Spam Automatically." To appear in the Proceedings of the *2006 Conference on Email and Anti-Spam* (CEAS'06), Mountain View, CA, July 2006.

[3] C. Pu, S. Webb, O. Kolesnikov, W. Lee, R. Lipton, "Towards the Integration of Diverse Spam Filtering Techniques." In the Proceedings of the *2006 IEEE International Conference on Granular Computing* (GrC'06), May 2006, Atlanta. Invited keynote presentation.

[4] C. Pu, S. Webb, S. Chitti, and J. Parekh . "A Case Study of Learning Filters in Spam Arms Race: Resistance to Camouflage Attacks." Submitted for publication.

[5] S. Webb and C. Pu. "Using Large Corpora for Spam Email Classification Experiments." Submitted for publication.

[6] G. Conti and M. Ahamad. "A Framework for Countering Denial-of-Information Attacks." *IEEE Security and Privacy*, Nov/Dec 2005.

[7] R. Perdisci, D. Dagon, W. Lee, P. Fogla, and M. Sharif, "Misleading Worm Signature Generators Using Deliberate Noise Injection". *IEEE Symposium on Security and Privacy*, 2006.

[8] Kang Li and Zhenyu Zhong, "Fast Statistical Spam Filter by Approximate Classifications", in *Proceedings of ACM SIGMETRICS 2006/IFIP Performance 2006*, July 2006.

[9] M. Srivatsa, L. Xiong and L. Liu, "TrustGuard: Countering Vulnerabilities in Reputation Management for Decentralized Overlay Networks." In the Proceedings of 2005 International Conference on World Wide Web", (WWW2005), May 10-14, 2005, in Chiba, Japan.

[10] Tom Jagatic, Nathaniel Johnson, Markus Jakobsson, and Filippo Menczer, "Social Phishing", to appear in *Communications of ACM*.